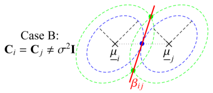
# Notation

| $p$ for continuous probabilities | $P$ for discrete probabilities |
|---|---|
| $H_0$ null Hypothesis | $H_1$ alternative Hypothesis |

$P_{ij} = P(\hat\omega = \omega_i, \omega = \omega_j)$
$\bar P_{ij} = P(\hat\omega = \omega_i \mid \omega = \omega_j)$
minimax : minimizes the maximum

# General stuff

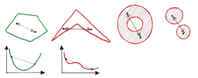
parameter estimation : continuous values vs DPR : discrete values
Mahalanobis distance : $D(\mathbf{x}, \mu) = \sqrt{(\mathbf{x}-\mu)^\top \mathbf{C}^{-1} (\mathbf{x}-\mu)}$
Euclidian distance : $D(\mathbf{x}, \mu) = \sqrt{(\mathbf{x}-\mu)^\top (\mathbf{x}-\mu)}$




## Matrix & Math basics

$(cA)^{-1} = c^{-1}A^{-1}$
$\det\left(A^{-1}\right) = (\det A)^{-1}$
$\partial X^{-1} = -X^{-1}(\partial X)X^{-1}$ (mit $XX^{-1} = I$ und $\partial(I) = 0$)
$\frac{\partial}{\partial x} x^T \mathbf{B} x = x^T(\mathbf{B} + \mathbf{B}^T)$
$\|\underline{x}\|^2 = \underline{x}^T \underline{x}$
$\log_b(P \cdot Q) = \log_b P + \log_b Q$
$\log_b P^n = n \cdot \log_b P$
$x_{1,2} = \dfrac{-b \pm \sqrt{b^2 - 4ac}}{2a}$
$(\frac{u}{v})' = \frac{u'v - uv'}{v^2}$

**Lagrange function :**
min / max$_{\underline{x}}$ $f(\underline{x})$ with equality constraints (s.t. $h_i(\underline{x}) = 0 \quad i \in [1, q]$) :
$L(\underline{x}, \underline{\lambda}) = f(x) + \sum_{i=0}^{q} \lambda_i h_i(\underline{x})$
with $\underline{\nabla}_{\underline{x},\underline{\lambda}} L(\underline{x}^*, \underline{\lambda}^*) = \underline{0}$ (stationary point)
+ with inequality constraints (s.t. $g_i(\underline{x}) \leqslant 0 \quad i \in [1, p]$) :
$L(\underline{x}, \underline{\alpha}, \underline{\lambda}) = f(\underline{x}) + \sum_{i=1}^{p} \alpha_i g_i(\underline{x}) + \sum_{i=1}^{q} \lambda_i h_i(\underline{x})$
with $\underline{\nabla}_{\underline{x},\underline{\alpha},\underline{\lambda}} L(\underline{x}^*, \underline{\alpha}^*, \underline{\lambda}^*) = \underline{0}$
mit $\underline{\alpha}$ und $\underline{\lambda}$ lagrange multipliers or dual variables
(Non)Convex :
for function no point above line between two points



minimax : min max maximin : max min
AWGN : additive white Gaussian noise

# Probabilities

(unknown) true state : $H$ ; estimated state : $\hat H$ ;

| $\hat H \mid H$ | $H = H_0$ | $H = H_1$ |
|---|---|---|
| $\hat H = H_0$ | true negative<br>$P_{\text{TN}} = 1 - P_{\text{FP}}$<br>specificity, correct rejection rate | false negative<br>$P_{\text{FN}} = 1 - P_{\text{TP}}$<br>miss rate |
| $\hat H = H_1$ | false positive<br>$P_{\text{FP}}$<br>false alarm rate | true positive<br>$P_{\text{TP}}$<br>sensitivity, hit rate<br>recall, detection rate |

## Bayes

$f(\theta \mid x) = \frac{f(x,\theta)}{f(x)} = \frac{f(x \mid \theta) f(\theta)}{f(x)}$ in our case :
$P(\omega_j) = \frac{p(\underline{x} \mid \omega_j) P(\omega_j)}{\sum_{k=1}^{c} p(\underline{x} \mid \omega_k) P(\omega_k)} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}$

## Error rate (ER)

Measures performance of classifier of major classes $\text{ER} = \sum\sum_{i \neq j} P_{ij}$

## Balanced error rate (BER)

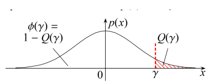
Measures performance of classifier with equal priors
$\text{BER} = \text{ER}\,|_{p(\omega_i) = \frac{1}{c}} = \sum\sum_{i \neq j} P(\hat\omega = \omega_i | \omega = \omega_j) \cdot P(\omega = \omega_j)$
$= \frac{1}{c} \sum\sum_{i \neq j} P(\hat\omega = \omega_i | \omega = \omega_j)$

## Gaussian/normal distribution

PDF : $p(\underline{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{|\mathbf{C}|}} e^{-\frac{1}{2}(\underline{x} - \underline{\mu})^T \mathbf{C}^{-1} (\underline{x} - \underline{\mu})}$ $d$ : #Features
bzw. $p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$
mit $\underline{x} \sim \mathcal{N}(\mu, \sigma^2)$ and $|\mathbf{C}| = \det(\mathbf{C})$ $\det(\sigma^2 \mathbf{I}) = \sigma^{2d}$
Mean $E(\underline{x}) = \underline{\mu} = \frac{1}{N} \sum_{n=1}^{N} x_n$
Covariance $\mathbf{C} = \mathrm{E}[(\underline{x} - \underline{\mu})(\underline{x} - \underline{\mu})^T] = \frac{1}{N} \sum_{n=1}^{N} (\underline{x_n} - \underline{\hat\mu})(\underline{x_n} - \underline{\hat\mu})^T$
Contour lines : ellipsoids centered at $\underline{\mu}$. Direction & size of principal axes given by eigenvectors of $\mathbf{C}$

## Standard Gaussian/normal :

PDF : $p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$
mit $\underline{x} \sim \mathcal{N}(0, 1)$




Q-function : $Q(x) = 1 - \int_{-\infty}^{x} p(z)\, dz$

# Loss functions

Loss-Matrix $[l_{ij}]$ with $l_{ij} = l(\hat\omega = \omega_i, \omega = \omega_j) \geqslant 0$

## 0/1-Loss

$l_{ij} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$ All errors are equally bad

# Bayesian decision

| | MBR | MAP (BBR) | ML (BBR) | Minimax | Neyman-Pearson |
|---|---|---|---|---|---|
| data $\underline{x}$ | known | known | known | known | known |
| likelihood $p(\underline{x}\|\omega_j)$ | known | known | known | known | known |
| prior $P(\omega_j)$ | known | predetermined | predetermined | unknown / varying | unknown / varying |
| loss $l_{ij}$ | known | known | predetermined | known | unknown / varying |

predetermined, known, unknown / varying

## Minimum Bayesian Risk (MBR)

Bayesian Risk : $\text{BR} = \sum_i \sum_j l_{ij} P_{ij} = E_{\hat\omega,\omega}(l(\hat\omega, \omega))$
$= \int_{\mathbb{R}^d} \sum_{j=0}^{1} l(\hat H(\underline{x}), H_j)\ p(\underline{x}, H_j)\, d\underline{x}$
$\hat\omega_{\text{MBR}}(\underline{x}) = \arg\min_{\hat\omega} R(\hat\omega \mid \underline{x})$
with $R(\hat\omega(\underline{x}) = \omega_i \mid \underline{x}) = \sum_{j=1}^{c} l_{ij} P(\omega_j \mid \underline{x})$
**Likelihood ratio test :** $\text{LR(T)} = \frac{p(\underline{x}|\omega_1)}{p(\underline{x}|\omega_0)} \lessgtr_{\omega_1}^{\omega_0} = \frac{l_{10} - l_{00}}{l_{01} - l_{11}} \cdot \frac{p(\omega_0)}{p(\omega_1)} = \gamma$
**Decision regions :** $R_0 = \{x \mid \text{LR}(x) < \gamma\}$, $R_1 = \{x \mid \text{LR}(x) > \gamma\}$

## MAP (maximum a posterior)

*special case of MBR with 0/1-Loss*
$\Rightarrow$ BR = ER *error rate*
$\text{BR} = \sum\sum_{i \neq j} P_{ij}$
$\hat\omega_{\text{MAP}} = \arg\max f_i(\underline{x})$ mit $f_i(\underline{x}) = \ln p(\underline{x}|\omega_i) + \ln P(\omega_i)$
**Likelihood ratio test :** $\frac{p(\underline{x}|\omega_1)}{p(\underline{x}|\omega_0)} \lessgtr_{\omega_1}^{\omega_0} = \frac{p(\omega_0)}{p(\omega_1)} = \gamma_{\text{MAP}}$

## ML (maximum likelihood)

*special case of MAP with equal priors* $P(\omega_i) = \frac{1}{c}$
balanced error rate : $\text{BER} = \frac{1}{c} \sum\sum_{i \neq j} P(\hat\omega = \omega_i | \omega = \omega_j) P(\omega_j)$
**Likelihood ratio test :** $\frac{p(\underline{x}|\omega_1)}{p(\underline{x}|\omega_0)} \lessgtr_{\omega_1}^{\omega_0} = 1 = \gamma_{\text{ML}}$

## Minimax

Minimal expected error for "pessimistic" (worst case) priors

## Neyman-Pearson decision (NP) or CFAR

$\text{LR} = \frac{p(\underline{x} \mid \omega_2)}{p(\underline{x} \mid \omega_1)} \lessgtr_{\omega_2}^{\omega_1} \gamma(\alpha)$
$\gamma$ determined by $P_{\text{FA}} = \int_{R_2(\gamma)} p(\underline{x}|\omega_1)\, d\underline{x} \overset{!}{=} \alpha$ bzw.
$P_{\text{FA}} = \int_{\hat H(\underline{x}) = H_1} p(\underline{x}|H_0)\, d\underline{x}$

# Classifier

## Bayes plug-in

Naive Bayes plug-in : all features are independent (unlikely -> rough approx.) → $p(\underline{x}|\omega_i) = \prod_{i=1}^{d} p(x_i|\omega_j\ ;\ \vartheta ij)$

## PDF estimation : parzen window

with **Gaussian Kernel** to estimate PDF :
$\Phi(\underline{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{1}{2}||x||^2}$ with $\mathcal{N}(0, \underline{I})$

## SVM (support vector machine)

based on bin decision. Maximizes margin to sample.
LDF $f(\underline{x}) = \underline{w}^T \underline{x} + w_0$
Convex optimization problem (KKT)
Multi-class via : one against one, hierarchical, (one against rest)

## Hard Margin SVM

Binary dataset (1..N samples) linearly separable (hard margin)
Primal problem : $\min_{\underline{\omega}, \omega_0} f(\underline{\omega}, \omega_0) = \frac{1}{2}||\underline{\omega}||^2$ s.t.
$g_n(\underline{\omega}, \omega_0) = 1 - y_n(\underline{\omega}^T x_n + \omega_0) \quad 1 \leqslant n \leqslant N$
⇒ convex quadratic problem → KKT sufficient, strong duality
Lagrange function : $L(\underline{\omega}, \omega_0, \underline{\alpha}) = \frac{1}{2}||\underline{\omega}||^2 + \sum_n \alpha_n g_n(\underline{\omega}, \omega_0)$
$= \frac{1}{2}||\underline{\omega}||^2 - \underline{\omega}^T \underline{\mathbf{X}}\underline{\alpha} - \omega_0 \underline{y}^T \underline{\alpha} + \underline{1}^T \underline{\alpha}$
with $\underline{\mathbf{X}} = [y_1 \underline{x}_1,\ y_2 \underline{x}_2,\ \dots,\ y_N \underline{x}_N]$
⇒ KKT conditions (4.5.4.3)
Use dual problem to determine $\alpha_n$ :
$d(\alpha) = -\frac{1}{2}\underline{\alpha}^T \mathbf{Q} \underline{\alpha} + \mathbf{1}^T \underline{\alpha}, \quad \mathbf{Q} = \mathbf{X}^T \mathbf{X}$
with $\alpha_n$ :
$\underline{\omega} = \mathbf{X}\underline{\alpha} = \sum_{n \in \mathrm{SV}} \alpha_n y_n \underline{x}_n, \qquad \omega_0 = \frac{1}{|\mathrm{SV}|} \sum_{n \in \mathrm{SV}} (y_n - \underline{\omega}^T \underline{x}_n)$
Support-Vectors : all $\underline{x}_n$ with $\underline{\alpha_n} > 0$ and must be on boundary ($g_n() = 0$, but only sufficient)
Only SV determien solution. Use Dual problem to find $\alpha_n$

## Soft Margin SVM

if not linearly seperable (even after feature mapping)
Slack variable $\xi_n$ : Optimization s.t. :
$y_n(\underline{\omega}^T x_n + \omega_0) \geqslant 1 + \xi_n$

# Decision tree

Binary decision tree, not robust
recursive splitting of dataset, can handle nominal (symbolic) features → nonmetric method.
metric impurity $H(S_n)$ e.g. Gini-index : $H(S_n) = 1 - \sum_{j=1}^{c} p_j^2$
Search for biggest $\Delta H$ reduction
Pruning : Cut subtrees with small impurity reduction (to avoid overfitting)

## RF (Random forest)

many decision trees
essemble method : combine models (with random samples of data/feature-set)

# Overview

| | kMeans | nearest mean | NN | SVM | kNN |
|---|---|---|---|---|---|
| supervised | 🟥 | 🟩 | 🟩 | 🟩 | 🟩 |
| multi-class | 🟩 | 🟩 | 🟩 | 🟥 | 🟩 |
| nonlinear dec. bound. | 🟥 | 🟥 | 🟩 | 🟥 | 🟩 |

# Discriminant functions

learn discriminant functions from training samples → no pfd required

## LDF (Linear discriminant functions)

decision boundaries are segments of hyperplanes $f_j(\underline{x}) = \underline{\omega}_j^T \underline{x} + \omega_{j0}$

# Features

## Feature mapping

e.g. polynomial mapping
$f_i(\underline{x}) = w_{j0} + \sum_{k=1}^{d} w_{j,k} x_k + \sum \sum_{k,l=1}^{d} w_{j,k,l} x_k x_l + \dots$

## Feature dimension reduction

to avoid overfitting : unknown parameters $\cdot 10 \ \approx\ N$ samples

## Feature selection

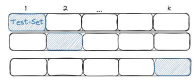
- exhaustive search (often not possible)
- single feature ranking (greedy algorithm - local min !)
- Heuristics : e.g SFS, SSFS (sequential forward selection)

## Feature transformation

LDA (linear discriminant analysis) uses Fisher ratio $F = \frac{|\sum_B|}{|\sum_W|}$

## k-cross validation

$k-1$ folds per training, with median ER




## Hyperparameter optimization

*e.g : type of classifier, kNN - number of k, GMM - number of models m, SVM - trade-off parameter C*
additional validation-set needed.

# KKT Condition

necessay conditions for local minimum :
- stationarity : $\underline{\nabla_x} L(\underline{x}^*, \underline{\alpha}^*, \underline{\lambda}^*) = 0$
- primal feasibility : $g_i(\underline{x}^*) \leqslant 0$ and $h_i(\underline{x}^*) = 0\ \forall i$
- dual feasibility : $\alpha_i^* \geqslant 0\ \forall i$
- complementary slackness : $\alpha_i^* g_i(\underline{x}^*) = 0\ \forall i$

## Primal and dual problem

primal problem $f^* = \min_{\underline{x}} f(\underline{x})$ s.t. $g_i(\underline{x}) \leqslant 0$ and $h_j(\underline{x}) = 0$ (with $i = [1, p]$, $j = [1, q]$)
can be converted to **dual problem :** $d^* = \max_{\underline{\alpha}, \underline{\lambda}} d(\underline{\alpha}, \underline{\lambda})$ s.t. $\alpha_i \geqslant 0$
$d(\underline{\alpha}, \underline{\lambda}) = \min_{\underline{x}} L(\underline{x}, \underline{\alpha}, \underline{\lambda})$
convex primal problems : strong duality : $f^* = d^*$ (in general $d^*$ only lower bound for $f^*$)